09/12/00 JC931 U.S. PTO

09-13-00

A

JC490 U.S. PTO 09/660025 09/12/00

IN THE UNITED STATES
PATENT AND TRADEMARK OFFICE

PATENT APPLICATION

Fabio M. Chiussi
Andrea Francini
Denis Andreyevich Khotimsky
Santosh Krishnan

CASE 21-9-8-5

TITLE Method And Apparatus Of Feedback Control In A Multi-Stage Switching System

ASSISTANT COMMISSIONER FOR PATENTS
WASHINGTON, D.C. 20231

SIR:

NEW APPLICATION UNDER 37 CFR § 1.53(b)

Enclosed are the following papers relating to the above-named application for patent:

Specification
10 Informal Sheets of drawing(s)
Declaration and Power of Attorney (Unsigned)

CLAIMS AS FILED				
	NO. FILED	NO. EXTRA	RATE	CALCULATIONS
Total Claims	16 - 20 =	0	x $18 =	$0
Independent Claims	1 - 3 =	0	x $78 =	$0
Multiple Dependent Claims, if applicable			+ $260 =	$0
Basic Fee				$690
			TOTAL FEE	$690

Please file the application and charge **Lucent Technologies Deposit Account No. 12-2325** the amount of $690, to cover the filing fee. Duplicate copies of this letter are enclosed. In the event of non-payment or improper payment of a required fee, the Commissioner is authorized to charge or to credit **Deposit Account No. 12-2325** as required to correct the error.

The Assistant Commissioner for Patents is hereby authorized to treat any concurrent or future reply, requiring a petition for extension of time under 37 CFR § 1.136 for its timely submission, as incorporating a petition for extension of time for the appropriate length of time if not submitted with the reply.

Please address all correspondence to **Docket Administrator (Room 3C-512), Lucent Technologies Inc., 600 Mountain Avenue, P.O. Box 636, Murray Hill, New Jersey 07974-0636**. However, telephone calls should be made to me at 732-949-6710.

Respectfully,

Jeffery J Brosemer
Reg. No. 36096
Attorney for Applicant(s)

"Express Mail" mailing label number EE821604825US
Date of Deposit Sept. 12, 2000
I hereby certify that this appl. is being deposited with the United States Postal Service "Express Mail Post Office to Addressee" service under 37 CFR 1.10 on the date indicated above and is addressed to the Commissioner of Patents and Trademarks Washington, D.C. 20231
KIM STONE
(Printed name of person mailing paper of fee)
(Signature of person mailing paper of fee)

Date: 9.12.00

Lucent Technologies Inc.
600 Mountain Avenue (Room 3C-512)
P.O. Box 636
Murray Hill, New Jersey 07974-0636

09660025.091200

METHOD AND APPARATUS OF FEEDBACK CONTROL IN A MULTI-STAGE SWITCHING SYSTEM

FIELD OF THE INVENTION

This invention relates generally to the filed of communication and switching systems and in particular, to a method of selective feedback control in multi-module multi-stage packet switches.

BACKGROUND OF THE INVENTION

As the demand for aggregate switching capacity increases, multiple-module switching systems are being adopted as a scalable alternative to a single-node output-buffered switch (see, for example, F.M. Chiussi, J.G. Kneuer, and Kumar V.P., "Low-cost scalable switching solutions for broadband networking: The ATLANTA architecture and chipset", IEEE Communications Magazine, 35(12):44--53, December 1997.) At the same time, a significant amount of effort has been spent in designing complex scheduling algorithms which provide diversified Quality of Service (QoS) guarantees such as bounded delay, guaranteed bandwidth and fairness to individual flows or virtual connections. However, most of these techniques regulate access to a single contention point, and hence are directly applicable to output-buffered switches and multiplexors. Such a model does not adequately represent many of the evolved switch architectures, which employ multiple stages such as the ingress port, switch fabric and egress port and have contention points associated with each stage.

Feedback control through the mechanism of *selective backpressure* is well known in the context of multi-stage switches. However, the prior work in this field has concentrated primarily on increasing the throughput of the switching system. In one such instance such as that described by F.M. Chiussi, Y.Xia and V.P.Kumar in an article entitled "Backpressure In Shared-Memory Based ATM Switches Under Multiplexed Bursty Sources", which apprered in Proc. IEEE INFOCOM' 96 – Networking the Next Generation, Volume 2, pp 830-843, San Francisco, CA, March 1996, whenever congestion occurs at an output link of the second stage, the first stage modules are pre-empted from sending traffic to that link by means of a per-output backpressure signal. Conceptually, it is possible to extend this idea by employing per-flow schedulers in each stage and using per-flow backpressure signals. However, such a replication of functionality in all the stages

defeats the purpose of building multi-stage switches, not to mention the increased implementation complexity. Hence, we are motivated to build a system in which we relegate all the fine grain scheduling details to the slower (in terms of aggregate switching capacity) first stage, aggregate a set of flows into a single session in the second stage, and provide an intelligent feedback mechanism that enables to maintain the QoS guarantees at the per-flow level. Some of the recent work done by D.C.Stephens and H.Zhang and presented as a paper entitled "Implementing Distributed Packet Fair Queueing in a Scalable Switch Architecture", which appeared in Proc. IEEE INFOCOM'98 – Gateway to the 21st Century, Volume 1, pp. 282-290, San Francisco, CA, March/April 1998, in multi-stage switches take a similar approach but do not address the issue of flow aggregation.

Employing queue length as an indication of congestion with asserting selective feedback with direct or indirect means is a well-known technique both in switching architectures and the Available Bit Rate (ABR) service. Use of the length of a fictitious, or virtual, queue instead of the length of an actual queue for that purpose has also been described before (See, for example, F.M.Chiussi, Y.Xia and V.P.Kumar, "Virtual Queueing Techniques for ABR Service: Improving ABR/VBR Interaction, Proc. IEEE INFOCOM'97 – Driving the Information Revolution, Volume 1, pp 406-418, Kobe, Japan, April 1997. The latter method provides a reliable indication of congestion caused by a specific traffic component (ABR type traffic) in the presence of other traffic components (Variable Bit Rate, or VBR, type traffic) when each traffic component is allocated its separate queue and the ABR queue is given strictly lower priority than the VBR queue. The length of a virtual ABR queue is controlled using the reference static ABR service rate and the derived actual arrival rate.

Docket No. 20400

SUMMARY OF THE INVENTION

The invention presents a method and an apparatus to assert selective feedback from a downstream stage to an upstream stage in a multi-stage switching system carrying traffic flows associated with multiple traffic components each of which is characterized by the specific Quality of Service (QoS) requirements. While the upstream stage of the system implements fine granularity queuing and a sophisticated scheduling discipline, the downstream stage allows the traffic flows of different components to be aggregated within the same queue associated with a specific pair of the input and output ports of the

downstream stage. The selective backpressure is asserted for the individual traffic components based on the value of the per-component credit function (credit counter) that is incremented with the service attributed by the downstream scheduler to the specific traffic component and is decremented on arrival of a packet belonging to that component from the upstream stage. If an aggregate downstream stage queue drains completely, the credit functions of all traffic components carried by that queue are reset to their respective maximum values. The per-component backpressure is asserted when the credit function reaches or falls below zero.

The disclosed credit-based backpressure mechanism provides a reliable means for per-class congestion indication in a multi-stage switching systems. It distinguishes itself from the prior art in that:

it uses the concept of service credit rather than actual or virtual queue length;

it provides an indication of congestion for multiple traffic components at the same time;

it controls the per-component credit counter using the actual services attributed to the given component by the downstream stage scheduler and the actual component arrival rate; it contains provisions to guarantee efficient redistribution of resources to the active traffic components in the case when one or more components underutilize their allocated share of resources; and

it employs a single aggregate queue for all traffic components in a downstream stage of the system, while relying on the upstream stage scheduler to control the composition of that aggregate queue.

09660265 - 062400

BRIEF DESCRIPTION OF THE DRAWINGS

The features and advantages of the present invention will become more readily apparent from the following detailed description of the invention in which:

FIG. 1 illustrates the structural components of a representative multi-stage system;

FIG. 2 illustrates a scheduling node in isolation;

FIG. 3 illustrates a representative backpressure mechanism based on observation of queue occupancy;

FIG. 4 illustrates an aggregation of flows in a common queue;

FIG. 5 illustrates a mechanism for generating credit-based feedback;

FIG. 6 illustrates a possible implementation of the credit-based feedback method in a two-stage switching system with cross-traffic flow aggregation;

FIG. 7 illustrates a possible implementation of the credit-based feedback method in a two-stage switching system with in-channel flow aggregation;

FIG. 8 illustrates an ingress port scheduling structure in a distributed scheduling architecture;

FIG. 9 illustrates a fabric scheduling structure in a distributed scheduling architecture; and

FIG. 10 illustrates a credit-based feedback mechanism for control of the Guaranteed bandwidth and Best Effort traffic components within a Non-Guaranteed Delay channel in a distributed scheduling architecture.

DETAILED DESCRIPTION OF THE INVENTION

With reference now to FIG.1, there is shown a general view of selected two adjacent stages in a multi-stage switching system. Upstream-stage queuing modules 10 are interconnected by internal links 30 to downstream-stage queuing modules 20. The system switches *flows*, i.e., abstract end-to-end traffic streams. Upstream scheduling nodes 15 and the downstream scheduling nodes 25 arbitrate access of the switched flows to the link bandwidth. Each of the upstream stage scheduling nodes 15 services a set of per-flow queues, which may be organized in a hierarchical structure, and feeds a set of the second stage queues and associated scheduling nodes. A *spatial channel* is defined by a pair of upstream and downstream scheduling nodes. Traffic belonging to a certain flow competes for service bandwidth with other flows of the same spatial channel (*in-channel* flows) and with traffic of other spatial channels sharing the same upstream and downstream scheduling nodes (*cross-traffic* flows). For a given spatial channel, the upstream cross-traffic flows and the downstream cross-traffic flows constitute two mutually exclusive sets. Therefore, the contention experienced at the upstream stage by the traffic in the channel is independent of the downstream contention experienced by the flows in the same channel.

A more detailed illustration of a scheduling node in isolation is presented in FIG.2. With reference to that FIG, scheduler 40 arbitrates access to the available capacity among a set 50 of sessions 52...58, which may correspond to individual traffic flows or their aggregates. Each session 52...58 is provided an individual queue 62...68, which may store the packets before they are scheduled for transmission. *Arrival rate* of a session is the dynamically varying rate at which its traffic enters the system. The session is *backlogged* if it has traffic waiting for service in the queue. The portion of available bandwidth

instantaneously provided to a session by a scheduler is referred to as *granted rate* of that session. The session's traffic leaves the node at the *output rate*. For an isolated scheduler, the output rate of each backlogged session is equal to its granted rate, unless restricted by an external *backpressure* signal. The selective backpressure signals 72...78 from the downstream portion of the system prevent specific session or group of session from being serviced by scheduler 40. Analogously, the selective backpressure signals 82...88 are generated for use by the upstream portion of the system. If backpressure is asserted for a specific session, the arrival rate of that session is reduced, most often to zero, within a fixed-length or variable length time interval known as backpressure latency.

In any single stage of a multistage system, scheduler 40 may implement an advanced scheduling algorithm, that provides high level Quality of Service (QoS) to its session. The QoS can be expressed in terms of bandwidth and/or delay guarantees as well as fairness. A scheduler in a packet switching system, which necessarily operates on a discrete data units, typically attempts to approximate an idealized fluid scheduling discipline used as a reference. A general scheduling reference is the Decoupled Generalized Processor Sharing (DGPS) approach, introduced in F.Toutain, "Decoupled Generalized Processor Sharing: A Fair Queueing Principle for Adaptive Multimedia Applications, Proc. IEEE INFOCOM'98 – Gateway to the 21st Century, Vol 1, pp. 291-298, San Francisco, CA March/April, 1998 and based on A.K.Parekh and R.G.Gallager, "A Generalized Processor Sharing Approach to Flow Control In Integrated Services Networks: The Single Node Case", IEEE/ACM Transactions on Networking, 1(3):344-357, June 1993. In DGPS each session is characterized by a *reserved*, or *guaranteed* rate g_i, the sum of guaranteed rates of all sessions being less than the service capacity, and excess bandwidth weight w_i. Let $W_i(t',t'')$ denote the amount of traffic of the i-th session served within time interval (t', t''), then for any two sessions i and j that are continuously backlogged during the interval (t', t''), the *decoupled fairness criterion* requires that:

$$\frac{W_i(t',t'')-g_i(t''-t')}{w_i}=\frac{W_j(t',t'')-g_j(t''-t')}{w_j}.$$

A fluid scheduler implementing DGPS is described by the following set of rules:

R1. Each backlogged session is granted at least its guaranteed rate g.

R2. For each backlogged session, the granted rate exceeds the guaranteed rate by a value proportional to weight w.

R3. For each non-backlogged session, the granted rate is equal to the lesser of its actual

arrival rate and the value determined according to rules *R1* and *R2*.

Referring to FIG.2, a DGPS scheduler that provides decoupled fairness criterion can be conveniently represented as containing two components: the Guaranteed Bandwidth Scheduling or GBS component 42 and the Excess Bandwidth Scheduling or EBS component 44. Both components operate in parallel, the GBS component being a non-work conserving while the EBS component being work-conserving. Whenever the GBS component selects a session, this service receives service. The session selected by the EBS component receives service only if the GBS component does not make a selection.

In a two-stage system of FIG.1, a feedback in the form of selective backpressure plays important role, when the traffic is downstream-constrained, *i.e.*, the granted rates at the upstream stage exceed those at the downstream stage. Assuming the stationary behavior of the system which is defined by the unchanged backlog status of all flows and the constant arrival rates of the flows which are not backlogged, the selective backpressure can be modeled using the proportional reduction of rate. The throughput R_i of a backlogged flow i can be found as

$$R_i = \min\left\{\beta_i r_i^{(1)}, r_i^{(2)}\right\},$$

where $r_i^{(k)}$ is the flow's granted rate at the k-th stage and β is the backpressure factor. Ideally, to ensure maximum throughput for the given flow while not imposing restrictions on the cross-traffic flow, backpressure allows to match the output rate of the upstream stage with the granted rate of the downstream stage:

$$\beta_i = \begin{cases} \dfrac{r_i^{(2)}}{r_i^{(1)}}, & \text{if } r_i^{(1)} > r_i^{(2)}; \\ 1, & \text{if } r_i^{(1)} \le r_i^{(2)}. \end{cases}$$

The commonly used method of backpressure generation is based on the observation of the queue occupancy, as shown in FIG.3. The feedback control block 90 contains a counter 91 and a comparator 92. The counter 91 is originally set to zero and subsequently incremented on event of a data unit arrival 95 to the queue 60 and decremented on the event of data unit departure 96 from the queue 60. The value of the counter 91, which is thus equal to the queue occupancy, is compared in the comparator 92 with threshold T which may be either or computed dynamically based on the state of the system. If the value of the counter 91 exceeds the threshold, a backpressure signal 97 is asserted towards the

upstream stage. The departure from the queue occurs when the queue is not empty, and the stage scheduler selects that queue for service. Referring to FIG.2, to ensure generation of the backpressure signals 82...88, a backpressure generation block 90 has to be replicated and attached to each of the queues 62...68.

The disclosed invention allows to reduce the number of queues in the downstream queuing node by aggregating queues 62...68 in a single queue multi-session FIFO queue 69, as shown in FIG. 4, provided that the individual service rates granted to each component session can be accounted for. Regardless of what component sessions within queue 69 is selected by the scheduler 40, the service is given to the data item at the head of that queue.

The method of the invention calls for assigning a value of credit function to one or more component session aggregated within the queue 69. Whenever the aggregate queue length reaches zero, the credit function of a component is initialized using a pre-computed static or state-dependent dynamic value. Subsequently, the credit function of a component is incremented when the scheduler selects that component for service, regardless of the component to which the data item at the head of the queue 69 actually belongs. The growth of the credit function is subject to truncation. The credit function of a component is decremented when a data item belonging to that component arrives to the queue. . For a given session k, the value of the credit function $C_k(t)$ is reset to its maximum value C_{k0} whenever the occupancy of the aggregate FIFO queue reaches and remains at zero. At other times, the credit change is governed by the following mathematical expression:

$$dC_k(t) = \begin{cases} \left(r_k^{(2)}(t) - a_k^{(2)}(t)\right) dt, & \text{if } C_k(t) < C_{k0}; \\ \left(r_k^{(2)}(t) - a_k^{(2)}(t)\right)^{-} dt, & \text{if } C_k(t) = C_{k0}. \end{cases}$$

Here $a_k^{(2)}(t)$ and $r_k^{(2)}(t)$ are the component second stage arrival and granted rates, respectively. The backpressure signal that blocks transmission of the k-th component flow from the upstream stage is asserted whenever $C_k(t) \leq 0$.

Under the assumptions that the backpressure signals are generated continuously and take effect instantaneously, the aggregate queue length is bounded by the sum of the

initial credits, and $C_k(t)$ never drops below zero. In practice, with finite backpressure sampling period and finite latency, the queue length remains bounded albeit by a larger value.

FIG. 5 shows the schematics of the backpressure generation engine implementing the disclosed feedback control method. For simplicity, it shows two component sessions, 52 and 54, aggregated in a single downstream queue 69. Whenever the scheduler 40 selects session 52 (signal 102) or session 54 (signal 104), a data item at the head of queue 69 departs the queuing stage. The feedback control block 100 contains a queue occupancy counter 105 with a comparator 106 and a set of credit modules 120, 140, one per each component session. The queue occupancy counter 105 is originally set to zero and subsequently incremented on event of a data unit arrival to the queue 69 and decremented on the event of data unit departure from the queue 69. When the queue occupancy is equal to zero, as determined by the comparator 106, a reset signal is provided to all credit modules. Each credit module 110, 120 contains a credit counter 121, 141 with a comparator 123, 143 and an initialization unit 122, 142. On receipt of a reset signal from comparator 106, the corresponding initial credit value provided by initialization units 122, 142, are written to credit counters 121, 141. While the value of the queue occupancy counter 105 differs from zero, the credit counters 121, 141 change values with data units arrivals from the upstream stage and services granted by the scheduler 40. For example, the credit counter 121, associated with session 52, is incremented whenever session selection signal 102 is generated by the scheduler 40, unless it is equal to the initial credit value supplied by initialization unit 122. The credit counter 121 is decremented when a data unit belonging to session 52 arrives to the queue 69 from the upstream stage. If the value of the credit counter 121 reaches or falls below zero, as determined by comparator 123, a backpressure signal for session 52 is asserted towards upstream stage. Credit module 140, associated with session 54, operates in a similar fashion. Nothing in this description prevents an implementation following this disclosure from using counters that operate in the reversed direction.

The disclosed credit-based feedback method ensures that if all sessions remain backlogged and downstream-constrained, the backpressure mechanism achieves per-flow credit equilibrium near its zero value with the arrival rates matching corresponding granted rates and provides the target throughput of $R_i = \min\{r_i^{(1)}, r_i^{(2)}\}$ with the optimal value of

the backpressure factor. If some of the flows are not backlogged or are downstream-constrained, then the service rate of the aggregate queue exceeds its stationary arrival rate and, therefore, the credit equilibrium can not be reached. The aggregate queue occupancy oscillates, periodically returning to zero and resetting the credit counters of all flows to their corresponding maximum values. Additionally, if the initial credit values for a set of sessions are statically selected in proportion to the excess bandwidth weight of the corresponding session than a common FIFO queue in a downstream stage that implements the disclosed credit-based feedback mechanism is sufficient to guarantee that the decoupled fairness criterion is met.

PREFERRED EMBODIMENTS

One embodiment of the disclosed invention allows to perform downstream aggregation of the spatial channels traversing different upstream queuing modules, as shown in FIG.6. For simplicity, each spatial channel is assumed to carry a single flow. Flows 211, 221, 231 traverse upstream queuing modules 210, 220, 230, respectively, and share downstream queuing module 250. Exemplified by upstream queuing module 210, the traffic of flow 211 which is configured as a session in the scheduling node 215 enters queue 212 serviced by that scheduler. Queues 213 and 214, also served scheduler 215, carry cross-traffic flows traversing downstream queuing modules other than 250. In the downstream queuing module 250, the traffic flows from the different upstream modules enter the same aggregate queue 252 served by scheduler 255. The feedback control block 258 monitors the state of the queue along with its arrival and departure events and maintains the credit function in the per-flow basis using the disclosed method. The backpressure signals from the feedback control block 258 are communicated to the upstream modules 210, 220, 230, where, when asserted, they prevent schedulers 215, 225, 235 from selecting the corresponding sessions for service.

Alternatively, the disclosed method is used to provide selective feedback for a set of in-channel flows sharing both upstream and downstream queuing modules, as shown in FIG.7. In an upstream queuing module 260, a set of flows belonging to the same spatial channel are queued individually in queues 261, 262, 263. A plurality of queues 264 represents upstream cross-traffic flows. Scheduling node 265 provide high-granularity service its sessions, including the flows of the given spatial channel. In the downstream

queuing module 270, the flows of the given spatial channel are aggregated in a common queue 271, which is served by the scheduler 275 along with other downstream cross-traffic flows, represented by plurality of queues 274. The feedback control block 278 monitors the state of the queue 271 together with its arrival and departure events according to the method disclosed in the foregoing presentation, and generates the per-flow backpressure signal to control the behavior of the upstream scheduler 265.

In another embodiment, the feedback method described in this disclosure can be used to provide inter-stage interaction in a distributed scheduling architecture. A reference distributed switch contains ingress port card stage with large amount of buffers, switching fabric with moderate-size buffers, and the bufferless egress port card stage. The per-flow queuing and sophisticated weighted fair scheduling is performed primarily in the ingress stage. For the purpose of QoS provisioning, all traffic flows are classified as either Guaranteed Delay (GD), Guaranteed Bandwidth (GB) or Best Effort (BE). The port card scheduling architecture, which has a hierarchical structure with the top layer represented by the GBS and EBS scheduling pair, is shown in FIG. 8. In the switching fabric, the flows of a specific QoS class belonging to a specific spatial channel are aggregated in a common QoS channel queue. The QoS channel queues destined to the same fabric output are served by the same output port EBS/GBS scheduling pair, as shown in FIG. 9. The non-hierarchical scheduling structure with fewer number of queues fits the moderate buffering capacity of the switching fabric. The further reduction in the number of fabric queues can be achieved by aggregating the traffic of the GB and BE QoS classes within the same fabric queue, referred to as the Non-Guaranteed Delay (NGD) channel queue.

The aggregation technique employs the disclosed method and is based on the observation of the queue length and the EBS granted rates in the port card and the fabric. As there is no direct correspondence between GB and BE QoS channels, on one hand, and the GBS and EBS scheduling components, on the other, a modification of the disclosed method is required. The queue NGD channel feedback is composed of the two signals. A queue-length-based backpressure signal is asserted towards the corresponding port card, when the occupancy of the NGD channel queue exceeds a threshold, which can be either static or dynamic. The queue-length-based backpressure signal prevents the port card scheduler from serving flows of either GB or BE QoS classes. A credit-based backpressure signal is asserted when the EBS credit function in the fabric reaches or falls below zero.

The credit function is set to the original value (pre-computed static or state-dependent dynamic) each time the occupancy of the NGD channel fabric queue reaches zero and is incremented, subject to truncation, on each EBS service that this queue receives. The credit increments may optionally be suppressed during the time intervals when the queue-length-based backpressure signal is asserted. The credit function is decremented each time a data unit served port card EBS scheduling component arrives to the fabric NGD channel queue. The arrival events may be optionally filtered (disqualified for the purpose of the credit function decrement) depending on the QoS class of the arriving item. FIG.10 depicts the feedback mechanism for a selected spatial channel. Within the port card 290, the flows belonging to the GB QoS channel are queued in a queue set 291, whereas the flows belonging to the BE QoS channel, in a queue set 292. These flows are served by the port card scheduler 295, along with the cross-traffic flows. The GB flow aggregate of the given spatial channel may be selected by both the GBS component 296 and the EBS component 297 of the scheduler 295. The BE flow aggregate of the given spatial channel may be selected by the EBS component 297 only. In the switching fabric 300, both GB and BE flows are aggregated in an NGD queue 302 of the given spatial channel. NGD queue 302 is configured as a session in the fabric scheduler 310, which contains two components: the GBS component 311 and the EBS component 312. The feedback control block 320 associated with the NGD queue 302 contains a queue occupancy counter 321 and credit counter 324. The queue occupancy counter, which is initially set to zero, is incremented on the event of data item arrival into the queue 302, and is decremented on the event of data item departure from the queue 302. The value of the counter 321 is compared in the comparator 323 with a threshold provided by the threshold computation module 322. If the value of the counter exceeds the threshold, a backpressure signal 331 is asserted. Backpressure signal 331 is used in the port card 290 to prevent the scheduler 295 from selecting both GB and BE flow aggregates of the given spatial channel. Referring again to the feedback control block 320 of the fabric 300, the value of the counter 321 is compared with zero in comparator 327 to generate a reset signal to the credit counter 324. On each such reset event, an initial credit value provided by module 328 is written into the credit counter 324. Subsequently, the value of the credit counter 324 is incremented each time the NGS queue 302 is serviced by the EBS component 312 of the fabric scheduler 310, and is decremented when a data item serviced by the EBS component 297 of the port card scheduler 295 enters the NGD queue 302. The credit counter 324 may optionally skip increment events during the time intervals when the backpressure signal 331 is asserted.

The credit counter 324 may optionally skip decrement events when the GB flow data item enters the NGD queue. The value of the credit counter 324 is compared with zero in comparator 325. Whenever the counter value is equal or less than zero, a backpressure signal 332 is generated. Backpressure signal 332 is used in the port card to prevent the scheduler 295 from selecting for service the BE flow aggregate of the given spatial channel.

Various additional modifications of this invention will occur to those skilled in the art and all deviations from the specific teachings of this specification that basically rely upon the principles and their equivalents through which the art has been advanced are properly considered within the scope of the invention as described and claimed.

0560025 - 050205

The invention claimed is:

1. A method regulating traffic in a communications network comprising the steps of:

aggregating one or more component traffic flows into a component traffic stream;

aggregating one or more component traffic streams into an aggregate stream;

carrying the aggregate stream in a single, FIFO queue; and

generating selective backpressure on selected ones of the component traffic streams such that selected ones of the component streams are desirably regulated.

2. The method according to claim 1 wherein said aggregation of the one or more traffic flows is performed according to the destination of the traffic flows and the similarity of the Quality of Service requirements of the traffic flows.

3. The method according to claim 3 wherein said aggregation of the one or more component traffic streams into an aggregate stream is performed according to the destination of the component traffic stream.

4. The method according to claim 3 wherein said aggregation is performed according to the absence of delay guarantees.

5. The method according to claim 4 wherein said aggregate stream is carried in a single, FIFO queue.

6. The method according to claim 5 wherein said generating selective backpressure step comprises the steps of:

maintaining an aggregate queue occupancy counter;

maintaining a credit counter for each component traffic stream; and

asserting selective backpressure for a specific one of the component traffic streams when the corresponding credit counter reaches a predetermined threshold.

7. The method according to claim 6 further comprising the steps of:

initializing the credit counter to a maximum value;

decrementing the counter when an item of specific type arrives in the aggregate queue;

incrementing the counter when the queue is given service granted to the specific type of traffic stream without regard to the type of data item which departs the single FIFO queue;

truncating the counter at a specific maximum level; and

resetting the counter to a maximum value when the occupancy of the aggregate queue falls to zero.

8. The method according to claim 6 wherein said backpressure asserting step is performed when the credit counter reaches as value of zero.

9. The method according to claim 4 wherein the two of said component traffic streams are the Guaranteed Bandwidth Traffic Stream and the Best Effort Traffic Stream, and wherein each data item arrival and departure event can be associated with either guaranteed or excess bandwidth service provided by a corresponding scheduler.

10. The method according to claim 9, wherein the generating selective backpressure step further comprises the steps of:

maintaining an aggregate queue occupancy counter;

maintaining a Best Effort credit counter;

asserting a first type of backpresure; and

asserting a second type of backpressure.

11. The method according to claim 10 wherein said first type of backpressure is applied towards both the Guaranteed Bandwidth Traffic Stream and the Best Effort Traffic Stream and wherein said second type of backpressure applies toward the Best Effort Traffic Stream.

12. The method according to claim 11 wherein said step of maintaining a Best Effort credit counter further comprises the steps of:

initializing the counter to a maximum value;

incrementing the counter when an excess bandwidth service is provided to said aggregate queue;

decrementing the counter when a data item arrival is associated with excess bandwidth service; and

resetting the counter to its maximum value each time th eoccupancy of said aggregate queue reaches a value of zero.

13. The method according to claim 12 wherein said incrementing step is not performed if the first type of backpressure is asserted.

14. The method according to claim 12, wherein said decrementing step is not performed if the arriving data item belongs to the Guaranteed Bandwidth Traffic Stream.

15. The method according to claim 10 wherein said step of asserting a first type of backpressure occurs whenever the aggregate queue occupancy conter exceeds a predefined threshold.

16. The method according to claim 10, wherein said step of asserting a second type of backpressure occurs whenever the Best Effort credit counter reaches a value of zero.

Attorney Docket No. 95-0526

ABSTRACT OF THE DISCLOSURE

The present invention focuses on the aggregation of flows belonging to different classes of non-guaranteed-delay traffic into a single FIFO queue in a downstream stage of the multi-stage switch. These include the guaranteed flows requiring bandwidth reservation, the best-effort flows that require a fair share of the excess bandwidth, and the flows that require both types of guarantee. We disclose a credit-based backpressure scheme which selectively controls the traffic originating from the previous stage of the system while achieving the goal of meeting the requirements of the individual flows. The credit function is maintained for each controlled traffic component in the aggregate session, and its objective is to balance the actual arrival rate of the component with the service rate dynamically granted by the downstream scheduler. The number of flows that can be aggregated is related to the complexity of maintaining the credit functions for the different traffic components.

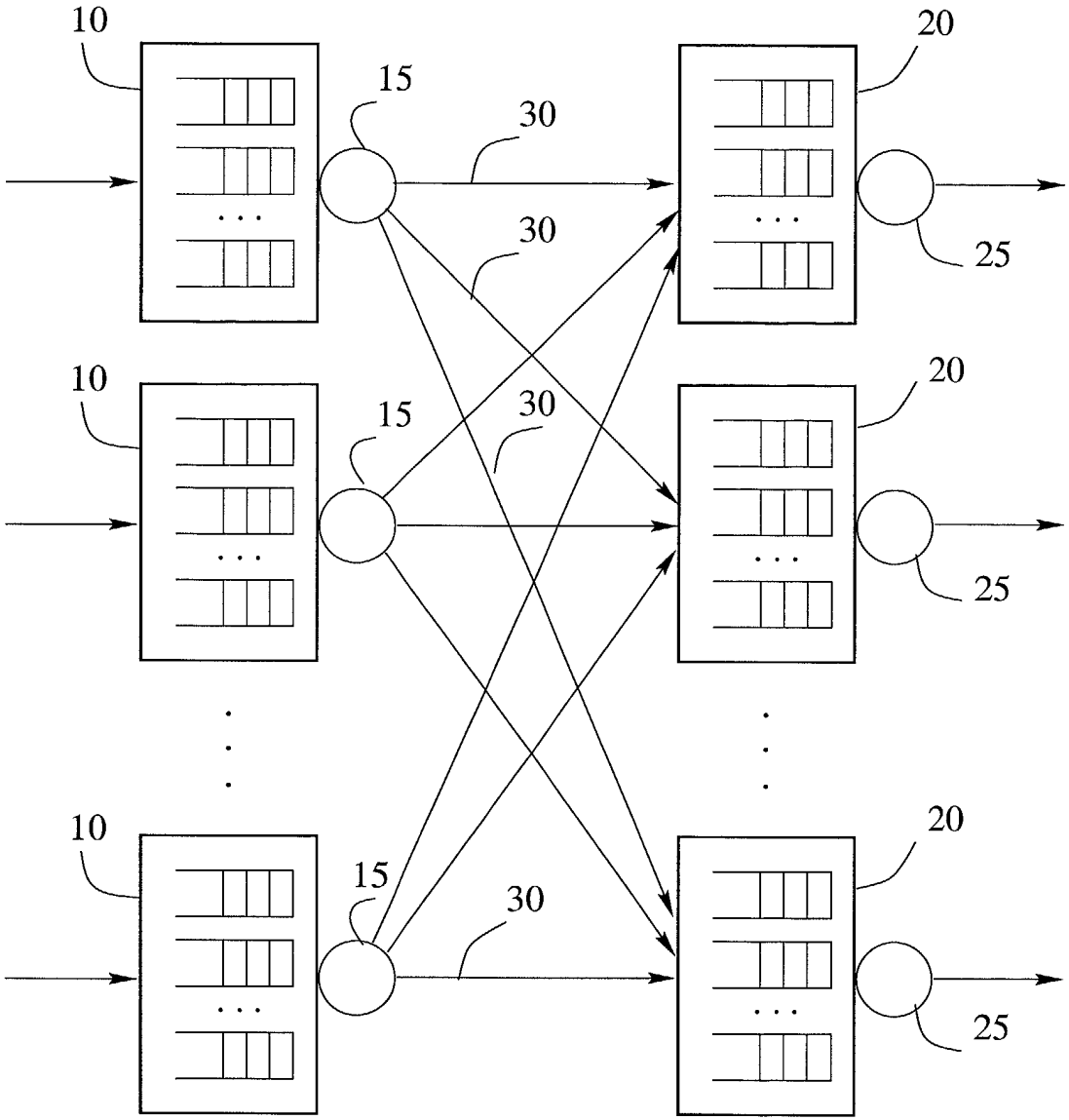

FIG. 1

09660025 .091200

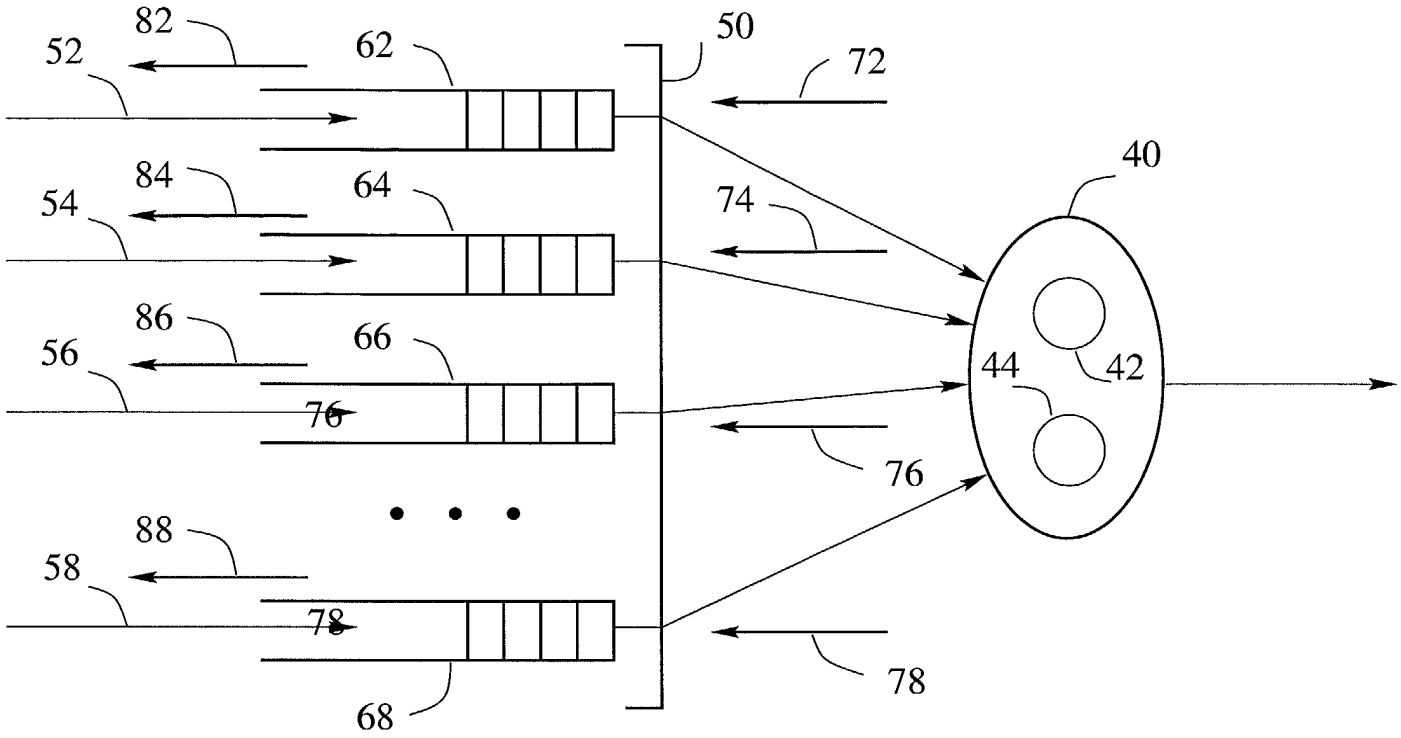

FIG. 2

09660025 .091200

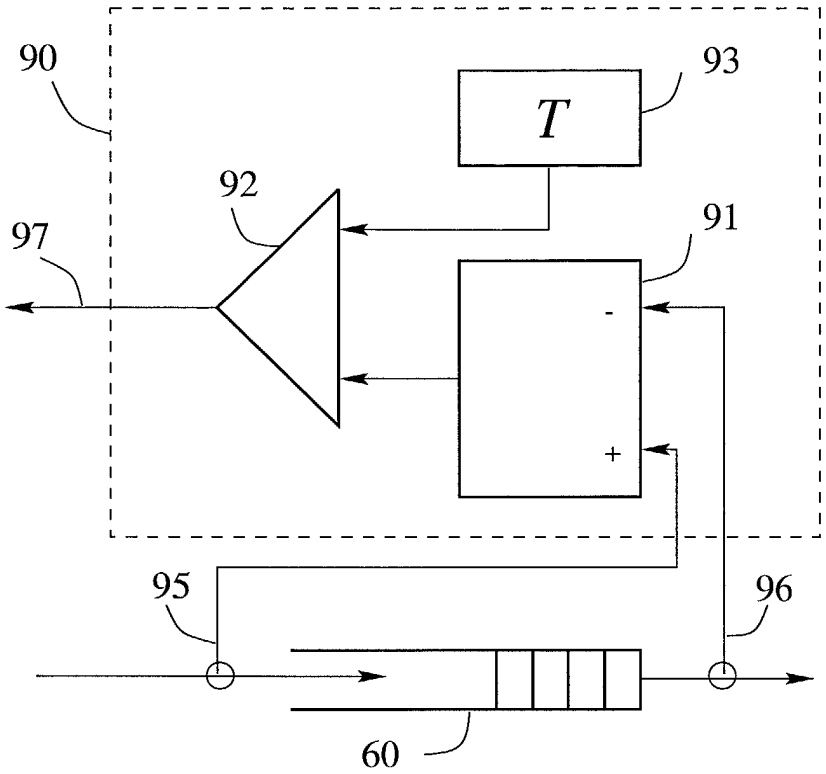

FIG. 3

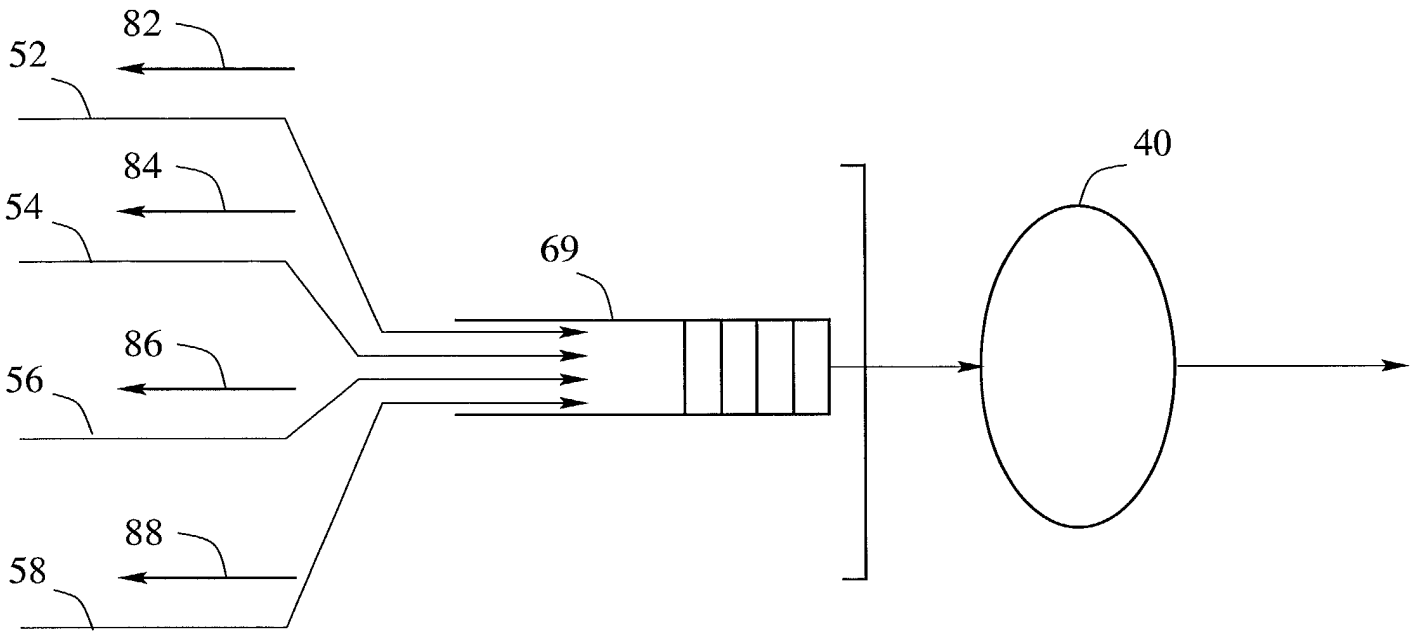

FIG. 4

09660025 . 091200

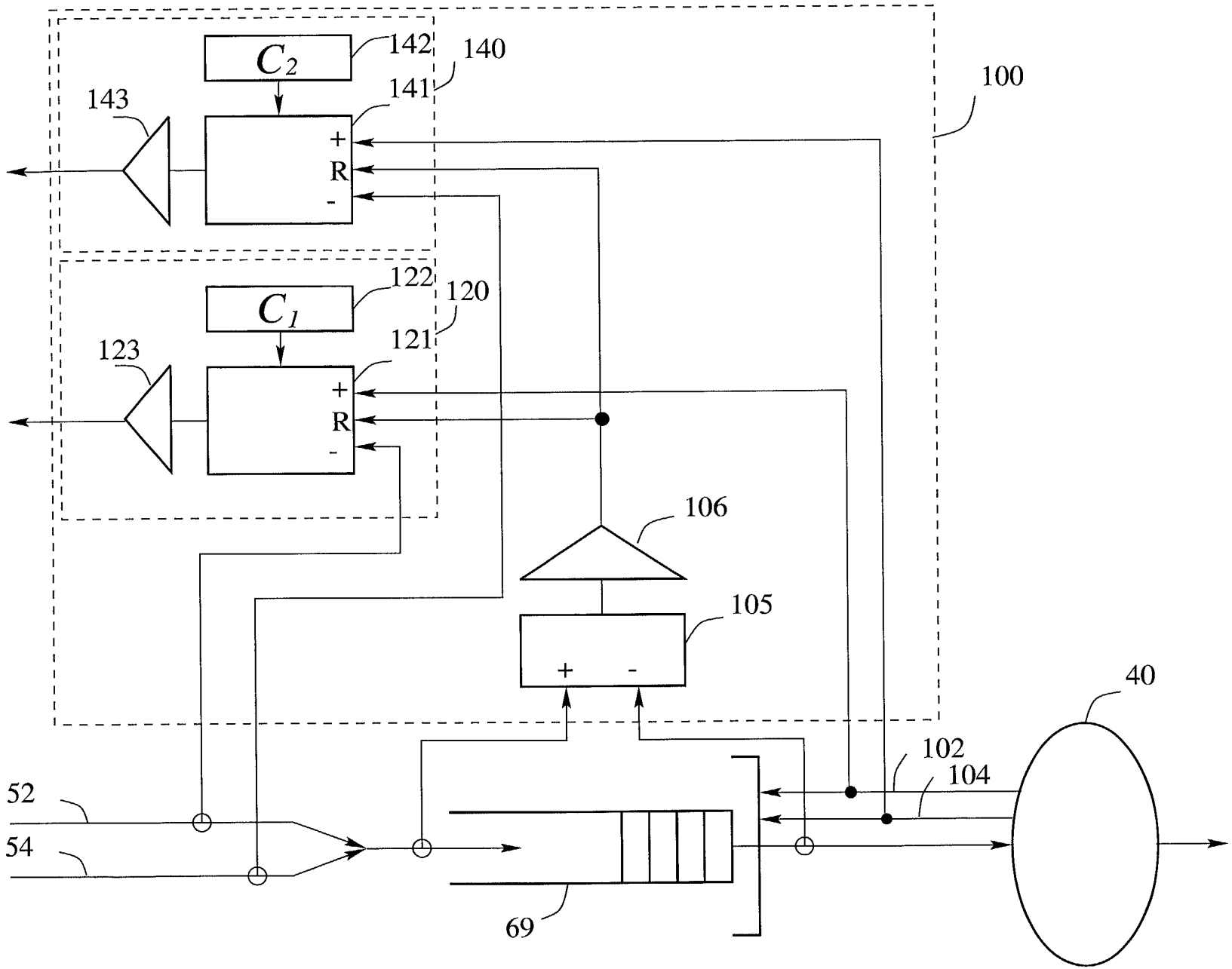

FIG. 5

09660025.091200

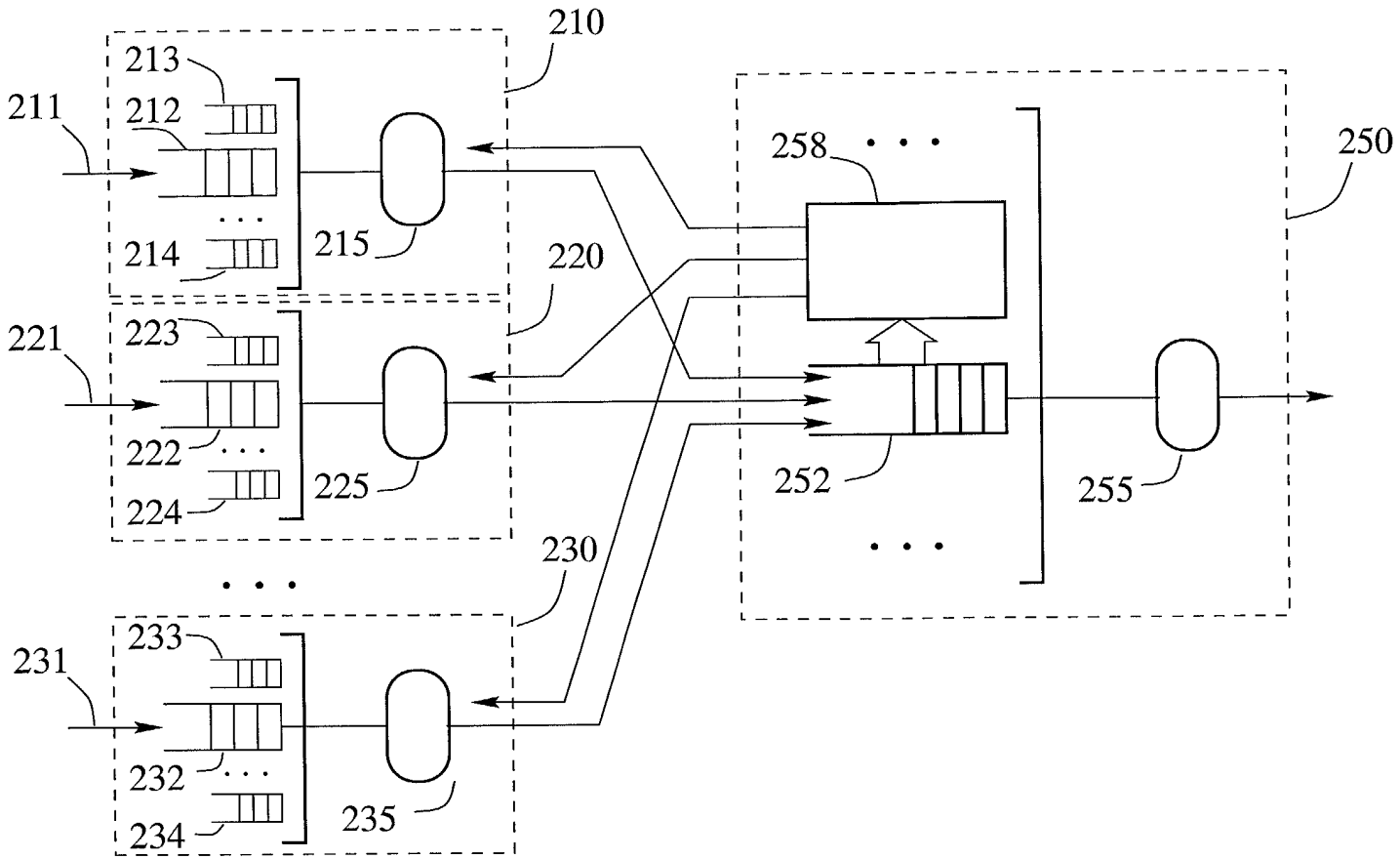

FIG. 6

09660025 .091200

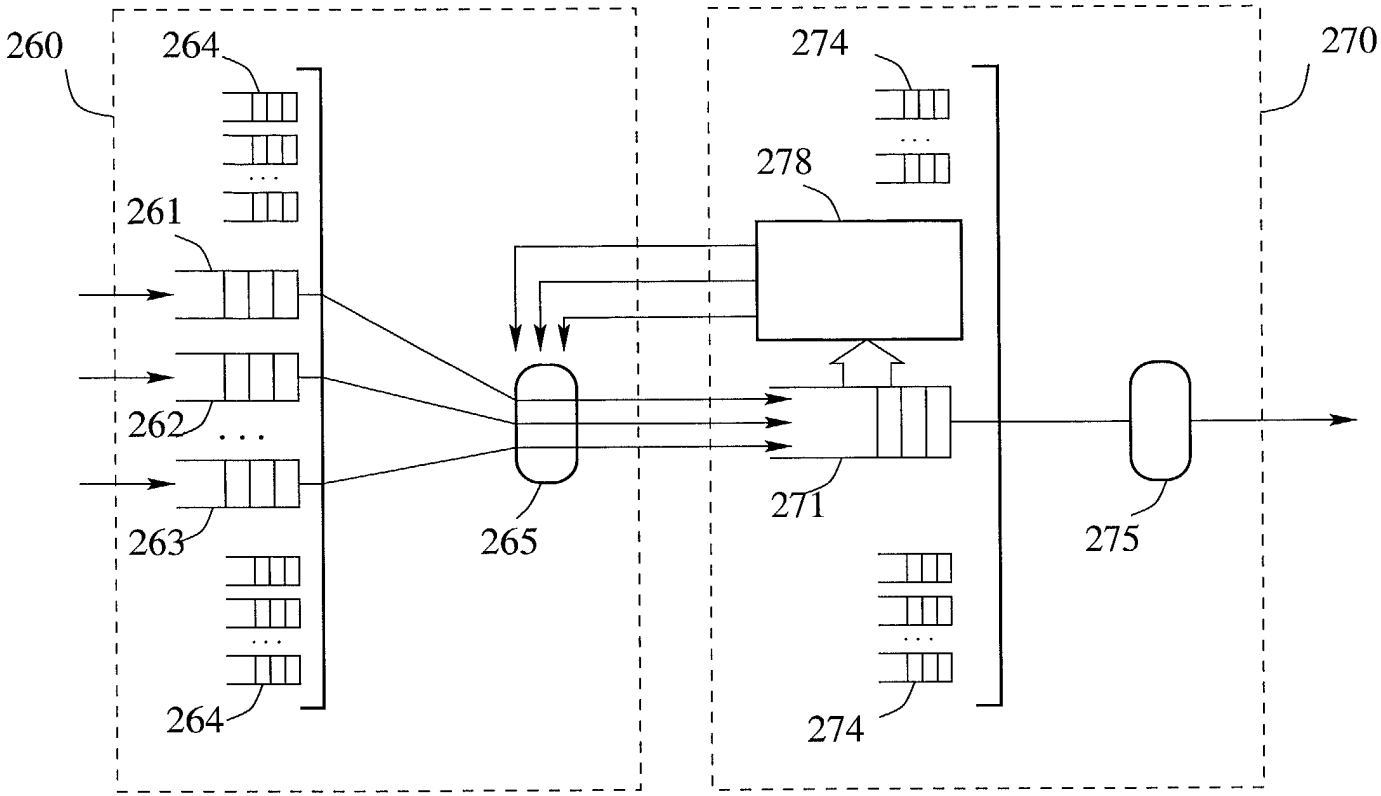

FIG. 7

09660025 .091200

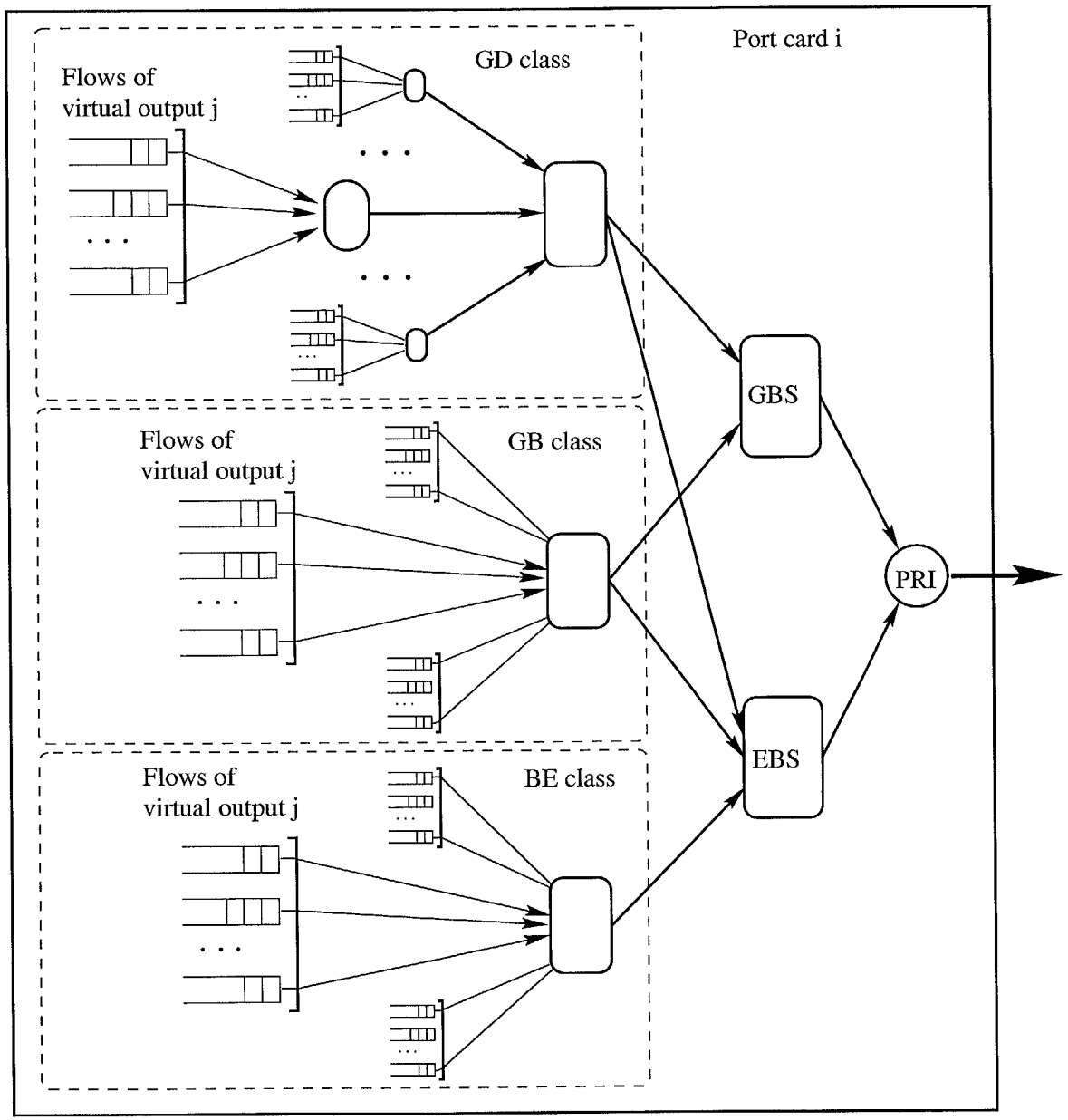

FIG. 8

09560025 . 091200

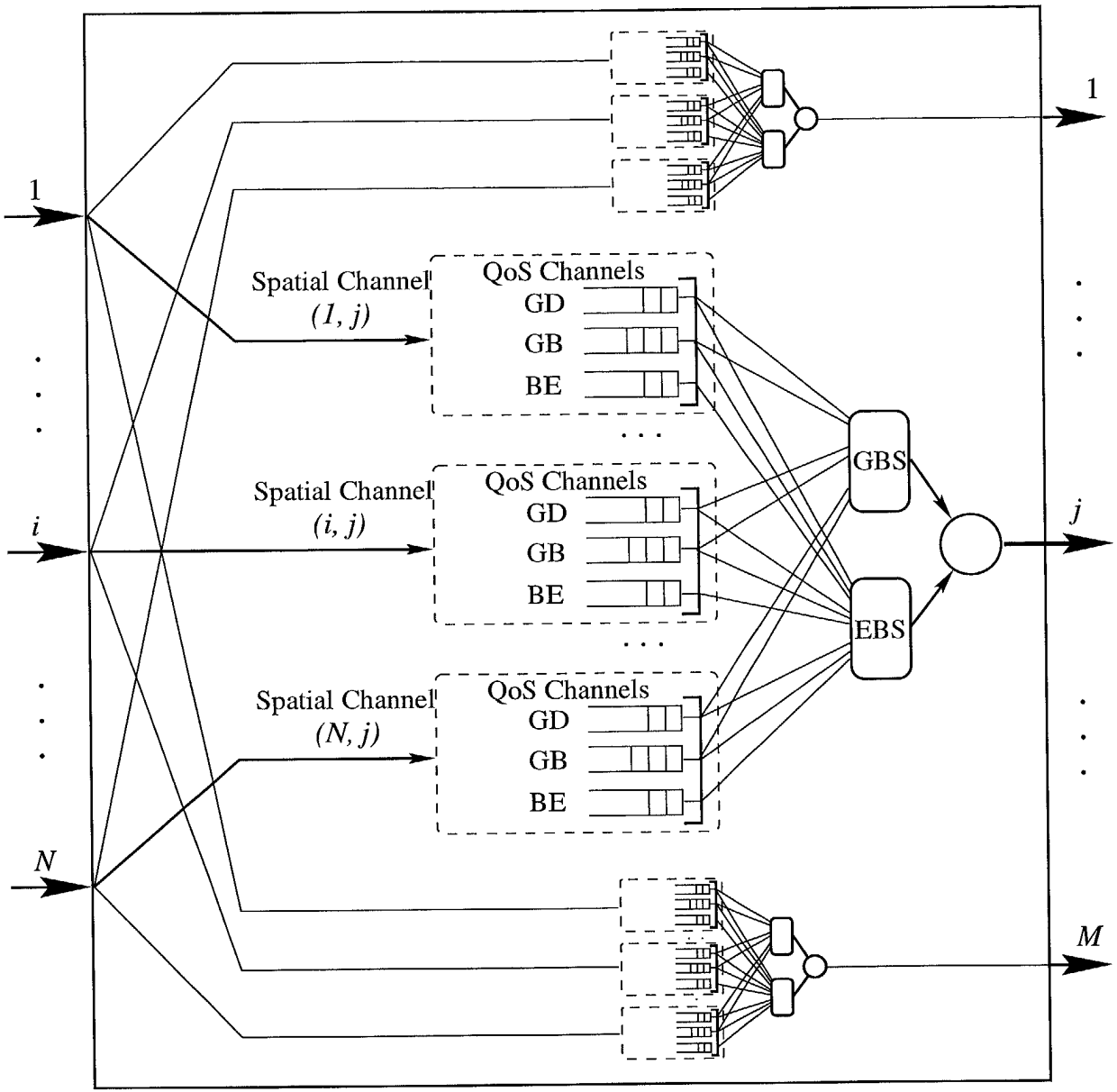

FIG. 9

09660025 . 091200

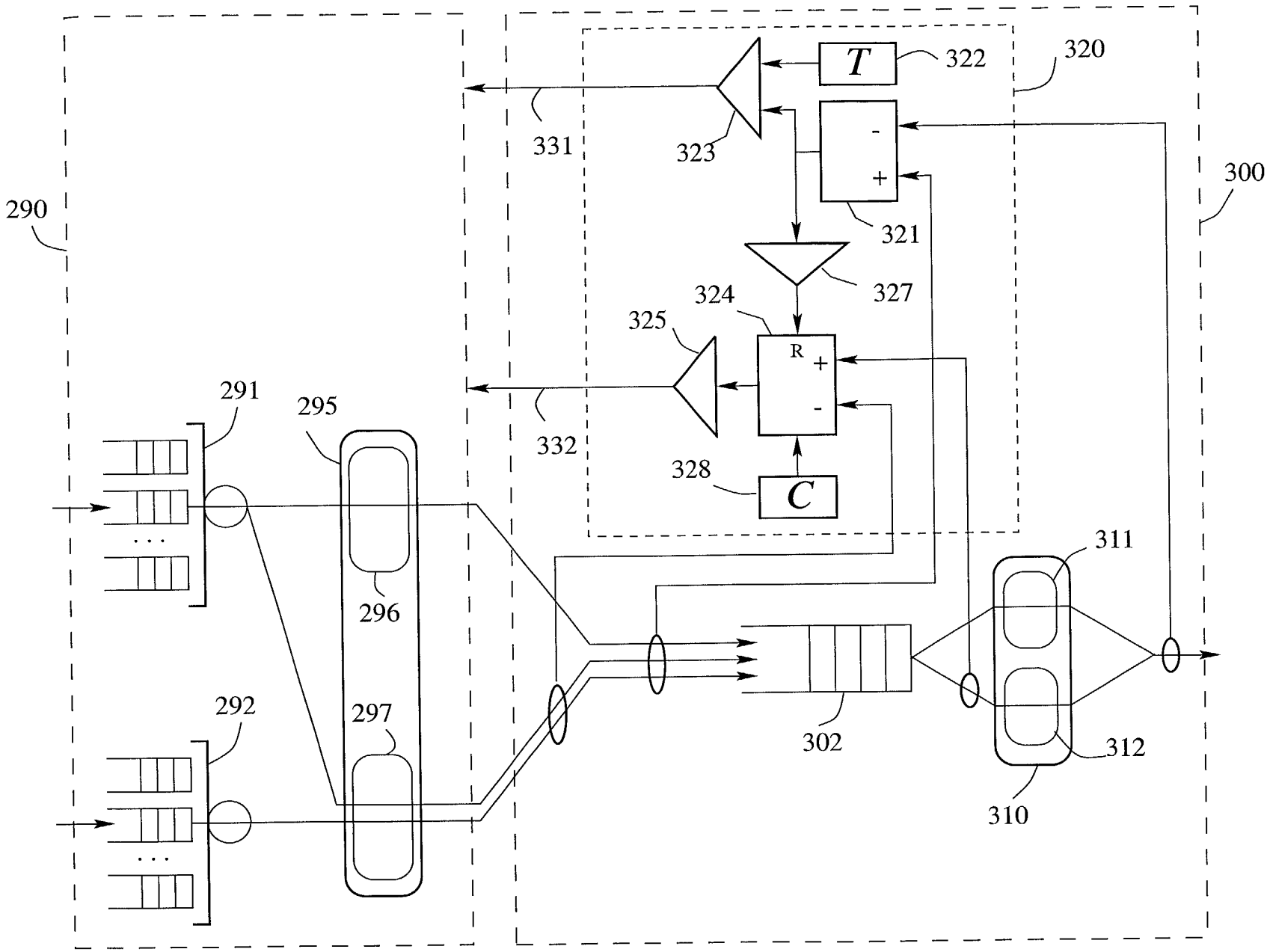

FIG. 10

09660025 .091200

IN THE UNITED STATES
PATENT AND TRADEMARK OFFICE

Declaration and Power of Attorney

As a below named inventor, I hereby declare that:

My residence, post office address and citizenship are as stated below next to my name.

I believe I am an original, first and joint inventor of the subject matter which is claimed and for which a patent is sought on the invention entitled **Method And Apparatus Of Feedback Control In A Multi-Stage Switching System** the specification of which is attached hereto.

I hereby state that I have reviewed and understand the contents of the above identified specification, including the claims, as amended by an amendment, if any, specifically referred to in this oath or declaration.

I acknowledge the duty to disclose all information known to me which is material to patentability as defined in Title 37, Code of Federal Regulations, 1.56.

I hereby claim foreign priority benefits under Title 35, United States Code, 119 of any foreign application(s) for patent or inventor's certificate listed below and have also identified below any foreign application for patent or inventor's certificate having a filing date before that of the application on which priority is claimed:

None

I hereby claim the benefit under Title 35, United States Code, 120 of any United States application(s) listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in the prior United States application in the manner provided by the first paragraph of Title 35, United States Code, 112, I acknowledge the duty to disclose all information known to me to be material to patentability as defined in Title 37, Code of Federal Regulations, 1.56 which became available between the filing date of the prior application and the national or PCT international filing date of this application:

None

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

I hereby appoint the following attorney(s) with full power of substitution and revocation, to prosecute said application, to make alterations and amendments therein, to receive the patent, and to transact all business in the Patent and Trademark Office connected therewith:

Thomas J. Bean	(Reg. No. 44528)
Lester H. Birnbaum	(Reg. No. 25830)
Richard J. Botos	(Reg. No. 32016)
Jeffery J. Brosemer	(Reg. No. 36096)
Kenneth M. Brown	(Reg. No. 37590)
Donald P. Dinella	(Reg. No. 39961)
Martin I. Finston	(Reg. No. 31613)
William S. Francos	(Reg. No. 38456)
Barry H. Freedman	(Reg. No. 26166)
Julio A. Garceran	(Reg. No. 37138)
Jimmy Goo	(Reg. No. 36528)
Anthony Grillo	(Reg. No. 36535)
Stephen M. Gurey	(Reg. No. 27336)
John M. Harman	(Reg. No. 38173)
Matthew J. Hodulik	(Reg. No. 36164)
Michael B. Johannesen	(Reg. No. 35557)
Mark A. Kurisko	(Reg. No. 38944)
Irena Lager	(Reg. No. 39260)
John B. MacIntyre	(Reg. No. 41170)
Christopher N. Malvone	(Reg. No. 34866)
John F. McCabe	(Reg. No. 42854)
Scott W. McLellan	(Reg. No. 30776)
Martin G. Meder	(Reg. No. 34674)
John C. Moran	(Reg. No. 30782)
Michael A. Morra	(Reg. No. 28975)
Gregory J. Murgia	(Reg. No. 41209)
Claude R. Narcisse	(Reg. No. 38979)
Joseph J. Opalach	(Reg. No. 36229)
Neil R. Ormos	(Reg. No. 35309)
Jack R. Penrod	(Reg. No. 31864)
Gregory C. Ranieri	(Reg. No. 29695)
Scott J. Rittman	(Reg. No. 39010)
Ferdinand M. Romano	(Reg. No. 32752)
Eugene J. Rosenthal	(Reg. No. 36658)
Bruce S. Schneider	(Reg. No. 27949)
Ronald D. Slusky	(Reg. No. 26585)
David L. Smith	(Reg. No. 30592)
Ozer M. N. Teitelbaum	(Reg. No. 36698)
John P. Veschi	(Reg. No. 39058)
David Volejnicek	(Reg. No. 29355)
Charles L. Warren	(Reg. No. 27407)
Eli Weiss	(Reg. No. 17765)

Please address all correspondence to the Docket Administrator (Rm. 3C-512), Lucent Technologies Inc., 600 Mountain Avenue, P. O. Box 636, Murray Hill, New Jersey 07974-0636. Telephone calls should be made to Jeffery J Brosemer by dialing 732-949-6710.

Full name of 1st inventor: Fabio M Chiussi

Inventor's signature _____ Date _____

Residence: Long Branch, Monmouth County, New Jersey

Citizenship: Italy

Post Office Address: 300 Ocean Avenue N., N4D
Long Branch, New Jersey, 07740

Full name of 2nd inventor: Andrea Francini

Inventor's signature _____ Date _____

Residence: Eatontown, Monmouth County, New Jersey

Citizenship: Italy

Post Office Address: 4 Victoria Drive
Eatontown, New Jersey, 07724

Full name of 3rd inventor: Denis Andreyevich Khotimsky

Inventor's signature _____ Date _____

Residence: Eatontown, Monmouth County, New Jersey

Citizenship: Russian Federation

Post Office Address: 42 Victoria Drive
Eatontown, New Jersey, 07724

Full name of 4th joint inventor: Santosh Krishnan

Inventor's signature _____ Date _____

Residence: Eatontown, Monmouth County, New Jersey

Citizenship: India

Post Office Address: 97C White Street
Eatontown, New Jersey, 07724

Docket No. 5050-1-3